

## Evolution of the let-7 microRNA Family

Jana Hertel,<sup>1</sup> Sebastian Bartschat,<sup>1,†</sup> Axel Wintsche,<sup>2,†</sup> Christian Otto,<sup>1,3,†</sup> The Students of the Bioinformatics Computer Lab 2011<sup>1</sup> and Peter F. Stadler<sup>1,4,5,6,7,8,\*</sup>

<sup>1</sup>Bioinformatics Group; Department of Computer Science and Interdisciplinary Center for Bioinformatics; University of Leipzig; Leipzig, Germany;

<sup>2</sup>Computational EvoDevo Group; Department of Computer Science and Interdisciplinary Center for Bioinformatics; University of Leipzig; Leipzig, Germany;

<sup>3</sup>LIFE - Leipzig Research Center for Civilization Diseases; Universität Leipzig; Leipzig, Germany; <sup>4</sup>Max Planck Institute for Mathematics in the Sciences;

Leipzig, Germany; <sup>5</sup>Fraunhofer Institut für Zelltherapie und Immunologie – IZI; Leipzig, Germany; <sup>6</sup>Department of Theoretical Chemistry; University of Vienna;

Wien, Austria; <sup>7</sup>Center for non-coding RNA in Technology and Health; University of Copenhagen; Frederiksberg C, Denmark; <sup>8</sup>Santa Fe Institute; Santa Fe, NM USA

<sup>†</sup>These authors contributed equally to this work.

The increase of bodyplan complexity in early bilaterian evolution is correlates with the advent and diversification of microRNAs. These small RNAs guide animal development by regulating temporal transitions in gene expression involved in cell fate choices and transitions between pluripotency and differentiation. One of the two known microRNAs whose origins date back before the bilaterian ancestor is mir-100. In Bilateria, it appears stably associated in polycistronic transcripts with let-7 and mir-125, two key regulators of development. In vertebrates, these three microRNA families have expanded to form a complex system of developmental regulators. In this contribution, we disentangle the evolutionary history of the let-7 locus, which was restructured independently in nematodes, platyhelminths, and deuterostomes. The foundation of a second let-7 locus in the common ancestor of vertebrates and urochordates predates the vertebrate-specific genome duplications, which then caused a rapid expansion of the let-7 family.

the class of microRNAs is subdivided into hundreds of families with presumably independent origins, the presence/absence of at least one representative of a family forms a set of valuable and phylogenetically highly informative characters.<sup>2,3</sup> While the loss of entire families is rare overall, some clades, such as in the tunicate *Oikopleura dioica*,<sup>4</sup> have undergone a major restructuring of their microRNA repertoire. On the other hand, individual microRNA families, with a few exceptions, evolve much like other multi-gene families showing gains by duplications, lineage-specific losses of paralogs, and reflect the genome-wide duplication events.<sup>3,5</sup>

Most microRNA families have only a few paralogous members making it fairly straightforward to resolve their evolutionary histories in full detail. The most complex example studied exhaustively to date is the mir-17 cluster, comprising about 15 microRNAs with eight different miRBase numbers that can belong to three unrelated families.<sup>6,7</sup> Apart from repeat-derived microRNAs and the huge imprinted mammalian microRNA clusters that behave in a repeat-like fashion, there is only a single ancient class of microRNAs whose evolutionary history has remained poorly understood: the let-7 family. This may come as a surprise, given that *Caenorhabditis elegans* let-7 and lin-4 were the first microRNAs to be discovered<sup>8</sup> more than a decade before microRNAs were recognized as a general class of RNA regulators,<sup>9</sup> and despite the fact that its phylogenetic distribution has been the

**Keywords:** microRNA, let-7, mir-100, mir-125, miRNA evolution

Submitted: 11/01/11

Revised: 11/30/11

Accepted: 12/08/11

<http://dx.doi.org/10.4161/rna.9.3.18974>

\*Correspondence to: Peter F. Stadler;  
Email: [studla@bioinf.uni-leipzig.de](mailto:studla@bioinf.uni-leipzig.de)

### Introduction

As a class, microRNAs exhibit several unusual features in their evolutionary history. Despite the short length of the functional mature microRNAs (miRs) with only ~22 nt, they are extremely well conserved and hence detectable with high accuracy in genomic sequence data.<sup>1</sup> Since

subject of systematic investigation already a decade ago.<sup>10,11</sup>

Both *let-7* and *lin-4* direct temporal development in different larval stages of *C. elegans*.<sup>8,12</sup> While *lin-4* miRNA does not seem to exist outside of Rhabditida (parasitic Roundworms), the *let-7* miRNA is conserved throughout Bilateria. Outside of Rhabditida, it is clustered with *mir-100* miRNA, in Panarthropoda and Deuterostomia this cluster is further extended by a *mir-125* as third microRNA, whose sequence is unrelated to both *let-7* and *mir-100*. Vertebrate genomes, on the other hand, typically contain a dozen or more *let-7* paralogs, some in clusters with paralogous copies of *mir-125* and/or *mir-100* and some located in isolation,<sup>13</sup> see also the data provided in miRBase Release 18.0.<sup>14</sup> Only part of this diversity can be accounted for by the genome-wide duplication events at the origin of vertebrates.<sup>15</sup> The major difficulty in deriving a comprehensive picture of the evolution of the *let-7* family is the correct assignment of orthology; unfortunately, the naming conventions used by miRBase are not helpful and even misleading in some cases. In addition, the annotated set of animal *let-7* sequences is still rather incomplete.

In this contribution, we therefore characterize the evolution of the *let-7* family and its associated microRNAs based on a comprehensive homology search and a careful analysis of their orthology relationships by means of both sequence comparison and assessment of synteny. We further suggest a refined nomenclature for the members of the *let-7* family that better reflects their evolutionary relationships.

## Materials and Methods

Starting point of our analysis was the collection of all *lin-4*, *let-7*, *mir-100*, and *mir-125* precursor sequences compiled in the miRBase database release 16. This includes 250 *let-7*, 86 *mir-125*, and 85 *mir-100* sequences in deuterostomes as well as 33 *let-7*, 26 *mir-125*, 27 *mir-100*, and 6 *lin-4* microRNAs in protostomes.

We performed a comprehensive homology search in the genomic sequences available for 60 deuterostomes, 50 protostomes, and five cnidarians. Therefore, we

extended all miRBase-derived microRNA-precursor sequences to a uniform length. Subsequently, BLAST<sup>16</sup> was applied with these sequences to search the genomes of 60 deuterostomes in order to collect a comprehensive data set. In cases where BLAST was not able to detect a certain homolog, we used the semi-global sequence alignment tool GotohScan<sup>17</sup> and finally Infernal.<sup>18</sup>

In order to firmly establish orthology relations, we determined for each microRNA gene its genomic context: for intergenic microRNAs, we recorded both adjacent protein-coding genes, for intronic locations, we recorded the surrounding genes. Homology of these protein-coding genes was established by alignments of the amino acid sequences whenever the corresponding information could not be retrieved from a database. Synteny among teleost species was determined from the pairwise alignment nets<sup>19</sup> provided through the UCSC genome browser. Regions with a size on the order of 100 kb were visually inspected for this purpose.

Taking into account the genomic locations, synteny information, and conservation of the 5p-miR/3p-miR regions, we built alignments for each microRNA family and its subfamilies. MicroRNA-like hairpin structures and secondary structure conservation were checked using RNAfold and RNAalifold from the Vienna RNA package.<sup>20,21</sup>

Analyses of deuterostome data were mainly based on manually curated alignments, calculated by ClustalW.<sup>22</sup> For phylogenetic analyses, we refined these alignments to a selection of taxa containing one member of primates (human), Laurasiatheria (dog), Metatheria (either wallaby or opossum), Protheria (platypus), Lepidosauria (anoles), Aves (either chicken or turkey), Amphibia (frog), and all available teleost sequences. Different members of the same cluster were aligned independently and then concatenated to increase the signal-to-noise ratio. SplitsTree<sup>23</sup> was applied for a visual investigation and refinement of our microRNA family assignment.

We estimated the phylogeny of the mixed clusters A, C, and D and the homogeneous clusters E–J using the program MrBayes v3.1.<sup>24</sup> Therein, the *mir-100* family of cluster C served as outgroup.

We used jModelTest<sup>25</sup> to select the best fitting nucleotide substitution model using the AIC. The JC and K80+I+G was selected for the mixed and homogeneous clusters, respectively. MrBayes was used to infer the posterior majority rule consensus tree along with posterior support for all internal branches using the respective evolutionary model and mainly default settings. The bayesian phylogenetic analyses of mixed and homogeneous clusters were run twice in parallel with eight (seven heated) Metropolis-coupled Markov chain Monte Carlo chains for 10 000 000 and 2 000 000 generations sampling every 1000th and 100th iteration, respectively. The initial 2 500 000 and 500 000 generations were discarded as burn-in during the estimation of the consensus trees of clusters A, C, and D and E–J, respectively. The trees are illustrated using Dendroscope.<sup>26</sup> All branches are labeled with their respective posterior support.

The Infernal package<sup>18</sup> was used for structure- and sequence-based homology searches whenever sequence conservation alone seemed to be insufficient. Therein, the program cmbuild was utilized with standard parameters to derive a CM for every *let-7*, *mir-100*, and *mir-125* (sub) family, given the corresponding cleaned alignment and its consensus secondary structure. In this scope, cleaned alignments solely contain full-length microRNA sequences that do not comprise any nucleotide except of A, C, G, or T. In addition, we created a compound *mir-100* CM based on a manually curated alignment of *mir-100* sequences derived from clusters A, C, and D for searches in Protostomia and basal Metazoa. Besides homology searches, these family models were used to determine the structural and sequence similarity of the *let-7* families in contrast to the phylogenetic analysis that was done with MrBayes. For each *let-7* family, we used all its microRNA sequences to calculate the average bitscore against each family CM.

## Results

Our survey resulted in a single copy of *let-7* in Protostomia, 14 *let-7* genes in human, and 19 *let-7* copies in teleosts, except for the zebrafish (*Danio rerio*)

where 21 genes could be retrieved. We note that mir-99 has long been known as a homolog of mir-100, while mir-98 is a let-7 homolog, see e.g. Roush and Slack.<sup>13</sup> In the following, we discuss the evolutionary history of the let-7 system in detail. As a resource, we provide extensive supplemental information in machine readable form, including *covariance models*, structure-annotated multiple sequence alignments, and the genomic coordinates of all microRNAs discussed in this work (<http://www.bioinf.uni-leipzig.de/publications/supplements/11-022>).<sup>1</sup>

**Let-7 microRNAs in basal Metazoans.** It is well known that miRNA mir-100 is one of the oldest miRNAs in animal species. The most ancient organism that has a mir-100 copy encoded in its genome is the sea anemone *Nematostella vectensis*, a cnidarian.<sup>27</sup> Somewhat surprisingly, no mir-100 ortholog was detectable in any of the diploplast genomes (cnidaria, ctenophorans, and poriferans), although this might be explainable by the incomplete status of these genome projects.

The let-7 microRNAs were detected by northern blot in a wide variety of both deuterostomes and protostomes, but remained undetectable in diploblasts.<sup>11</sup> Consistent results were obtained by microRNA sequencing.<sup>28</sup> Chaetognatha, which sometimes have been hypothesized as pre-dating the protostome-deuterostome divergence,<sup>29</sup> also have a let-7 homolog.<sup>11</sup> More recent phylogenetic studies, however, place them firmly within Protostomia.<sup>30,31</sup> No trace of a let-7 or mir-125 homolog can be found in any of the non-bilaterian animal genomes.

**Let-7 in Protostomes.** In most major protostome clades, we find a single intact cluster of mir-100, let-7, and mir-125. Typically, the cluster is tightly linked indicating an intact polycistronic transcript. This is the case for both lophotrochozoans and ecdysozoans with some exceptions.

Among lophotrochozoans, complete and tightly linked clusters are found in the annelids *Capitella teleta* and *Platynereis dumerilii*. In the latter, the expression of the let-7 cluster is studied in detail.<sup>32</sup> In the mollusc *Lottia gigantea*, the mir-125 homolog is missing. In contrast, the cluster has desintegrated in platyhelminthes and

mir-100 appears to be missing completely. Schistosomes have a single copy of let-7 and two mir-125 paralogs.<sup>33-35</sup> In *Schmidtia mediterranea*, multiple copies of let-7 and mir-125 as well as a single copy of lin-4 have been annotated.<sup>36-38</sup> In this scope, lin-4 can be seen as a putative homolog of mir-125. Both microRNA families show perfect conservation of their seed sequences, i.e., either nucleotides 1–7 or 2–8 of the 5p-miR region, compared with human let-7 and mir-125 miRs, respectively. Several substitutions are encountered in the remaining part of the 5p-miR sequences, however. Hence, the assignment of platyhelminth let-7 and mir-125 paralogs to particular subfamilies remains inconclusive.<sup>39</sup>

Much more genomic data are available for Ecdysozoa. In arthropods, mir-100, let-7, and mir-125 form a tight genomic cluster in which the microRNAs are separated only by a few hundred nucleotides. In *Drosophila melanogaster*, the polycistronic primary transcript and its expression has been studied in detail.<sup>40</sup> In a few species, one of the cluster members is lacking, possibly due to missing data. In nematodes, an intact cluster is present only in *Trichinella spiralis*, i.e., in the most basal clade Dorylaimia. In contrast, most rhabditid worms including *Caenorhabditis elegans* have an isolated let-7 gene and lack annotated mir-125 and mir-100 homologs. The loss of mir-100 appears to be a relatively recent phenomenon in *Caenorhabditis* and *Pristionchus*,<sup>41</sup> since a mir-100 linked to let-7 can be found in *Heterorhabditis bacteriophora*. Ruby et al. proposed, based on a match of the seed sequence, that the two related microRNA clusters mir-51/mir-53 (Chr. IV) and mir-54/mir-55/mir-56 (Chr.X) are co-orthologs of miR-100 in *C. elegans*.<sup>42</sup> The cluster on the Chr.X is separated from cel-let-7 by more than 1.5 Mb. Beyond the seed nucleotides, no homology with mir-100 is detectable, however, so that their relation with mir-100 remains uncertain.

Clusters comprising mir-100 and let-7 are also found in Tylenchina (e.g., Meloidogyne, Heterodera) as well as Spirurina (*Brugia malayi* and *Ascaris suum*). Poole and colleagues reported four mir-100 paralogs in *B. malayi*,<sup>43</sup> only one

of which, bma-mir-100b, is linked with the sole annotated let-7 and, furthermore, shows perfect sequence conservation with the human miR-100. The remaining three microRNAs show a conserved seed region but comprise various mutations in the 3' end of their 5p-miR sequence. None of these genomes contain a mir-125.

Lin-4, one of the first microRNAs to be discovered,<sup>44</sup> is functionally closely associated with let-7.<sup>45</sup> It was recognized as a putative ortholog of mir-125 by Lagos-Quintana et al.<sup>46</sup> based on the similarities of the 5p-miR regions. We find that the sequence homology covers the majority of the precursor hairpin supporting the homology of lin-4 and mir-125. In contrast to mir-125, however, none of the annotated lin-4 sequences is linked with let-7 and/or mir-100. *C. elegans* lin-4 is located in intron 9 of the protein-coding gene F59G1.4. This arrangement is conserved in both *Pristionchus pacificus* and *B. malayi*. No lin-4 sequence is detectable in *T. spiralis*, which has an intact mir-100/let-7/mir-125 cluster.

In *C. elegans*, there is an antisense transcript of lin-4, which could also give rise to a miRNA similar to the iab-4/iab-8 pair in *Drosophila*.<sup>47</sup> Thus, we checked whether lin-4 might originate from a mir-125 antisense hairpin. Comparisons of a lin-4 CM against annotated mir-125 sequences and a mir-125 CM against lin-4 sequences in both reading directions show that mir-125 and lin-4 match significantly better in sense direction. We thus hypothesize that the sequence divergence of lin-4 is coupled with the breaking up of the ancestral cluster.

**Let-7 in Gnathostomes.** In total, we collected 874 microRNAs among Deuterostomia including 128 mir-100 sequences, 135 mir-125 sequences, and 611 let-7 sequences. Most of these sequences were found in Gnathostomata (jawed vertebrates). The miRBase lists 12 let-7 paralogs in human including three genomic loci at which let-7 appears to be accompanied by other microRNAs. The best-known of these clusters, A, is composed of mir-99b, let-7e, and mir-125a on chromosome 19. The two other loci are C: mir-99a, let-7c, mir-125b-2 (chr.21) and D: mir-100, let-7a-2, mir-125b-1 (chr.11). The association

of mir-125 and let-7 at the latter two loci, although previously noticed, e.g., Roush and Slack,<sup>13</sup> is not annotated as a cluster in miRBase, since the distances of 50 and 46 kb, resp., are larger than the (arbitrary) 10kb threshold. The second type of let-7 cluster consists of members of the let-7 family only. The paradigmatic example is the cluster E on chromosome 9, consisting of let-7a-1, let-7f-1, and let-7d. The two remaining clustered loci are F on chr.X (let-7f-2 and mir-98) and G on chr.22 (let-7a-3 and let-7b). Two additional loci, designated here as I (chr.12) and J (chr.3), each harbour a single annotated human let-7 miRNAs (let-7i and let-7g, resp.). Our homology searches revealed two additional sequences similar to let-7d, located at positions K (chr.17) and L (chr.1), i.e., unrelated to the previously described let-7 loci.

With the exception of the novel loci K and L (see below), this arrangement of let-7 paralogs is well conserved: The three mixed clusters A, C, and D, the three homogeneous clusters E, F, and G, and the two isolated loci I and J can be traced throughout all available tetrapods. A summary of these gnathostome let-7 clusters is compiled in Table 1, the corresponding gene phylogenies are shown in Figure 1. An extended table showing annotated miRBase names and distances between adjacent microRNAs is also available online.

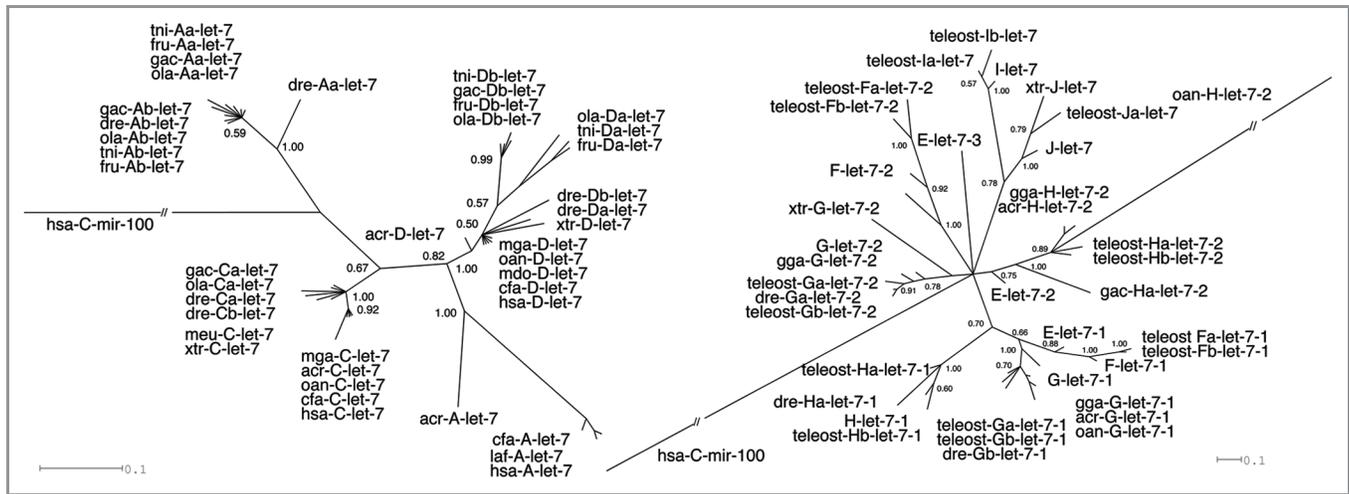
Orthology of the corresponding loci is unambiguously established based on both synteny information and sequence similarity (see Methods for details). In the chicken genome, two additional clustered let-7 paralogs, let-7k and let-7j, were reported.<sup>48</sup> They clearly form a fourth homogeneous cluster, H, absent in eutheria and metatheria. Evidence for the presence of the D, E, G, H, and J loci can also be found in the genome of the elephant shark. Since this genome is sequenced only at low coverage,<sup>49</sup> it is plausible that missing loci are due to lack of data rather than true losses.

The genomes of nearly all vertebrates, more precisely of the gnathostomes to the exclusion of lampreys and hagfishes, share two rounds of genome duplications.<sup>50-52</sup>

**Table 1.** Overview of miRNA clusters among Gnathostoma ordered by their presumed evolutionary history

| Org. | A-mir100<br>A-let-7<br>A-mir125 | C-mir100<br>C-let-7<br>C-mir125 | D-mir100<br>D-let-7<br>D-mir125 | E-let-7-1<br>E-let-7-2<br>E-let-7-3 | F-let-7-1<br>F-let-7-2 | G-let-7<br>G-let-7-1<br>G-let-7-2 | H-let-7<br>H-let-7-1<br>H-let-7-2 | K-let-7 | L-let-7 |
|------|---------------------------------|---------------------------------|---------------------------------|-------------------------------------|------------------------|-----------------------------------|-----------------------------------|---------|---------|
|      | A                               | C                               | D                               | E                                   | F                      | G                                 | H                                 | K       | L       |
| 1    | hsa                             | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 | ○       | ○       |
|      | ptr                             | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 | ○       | ○       |
|      | ggo                             | ○                               | ●●●                             | ●●●                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | ppy                             | ●●                              | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 | ○       | ○       |
|      | pha                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | mml                             | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 | ○       | ○       |
|      | cja                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | tsy                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | oga                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | mmr                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
|      | tbe                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 | ○       | ○       |
| 2    | mo                              | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 |         |         |
|      | mmu                             | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 |         |         |
|      | dor                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | str                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | cpo                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | opr                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | ocu                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
| 3    | fca                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | cfa                             | ●●●                             | ●●●                             | ●●●                                 | ○●○                    | ○●○                               | ●                                 |         |         |
|      | aml                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | bta                             | ●●●                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●                                 |         |         |
|      | oar                             | ○                               | ○                               | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | ssc                             | ●●●                             | ●●●                             | ●●●                                 | ○●○                    | ●●                                | ●                                 |         |         |
|      | vpa                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○                                 | ○                                 |         |         |
|      | ttr                             | ○●○                             | ○                               | ○●○                                 | ○●○                    | ○                                 | ○                                 |         |         |
|      | eca                             | ●●●                             | ●●●                             | ●●●                                 | ○●○                    | ○●○                               | ●                                 |         |         |
|      | mlu                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | pva                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | eeu                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | sar                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
| 4    | laf                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | ete                             | ○●○                             | ○                               | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
| 5    | pea                             | ○●○                             | ○●○                             | ○●○                                 | ○                      | ○●○                               | ○                                 |         |         |
|      | cho                             | ○●○                             | ○●○                             | ○●○                                 | ○                      | ○●○                               | ○                                 |         |         |
|      | dno                             | ○●○                             | ○●○                             | ○                                   | ○●○                    | ○●○                               | ○                                 |         |         |
| 6    | meu                             | ○                               | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○                                 |         |         |
|      | mdo                             | ○                               | ●                               | ●●●                                 | ●●●                    | ●●●                               | ●                                 |         |         |
|      | oan                             | ○                               | ●                               | ●●●                                 | ●●●                    | ●●●                               | ●●○                               |         |         |
| 7    | acr                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | tgu                             | ○●○                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●●●                               |         |         |
|      | apl                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | mga                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | gga                             | ○●○                             | ●●●                             | ●●●                                 | ●●●                    | ●●●                               | ●●●                               |         |         |
|      | xtr                             | ●                               | ●●●                             | ●●●                                 | ●                      | ●●●                               | ●                                 |         |         |
| 8    | tni                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | fru                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | gac                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | ola                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | dre                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | tni                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | fru                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | gac                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | ola                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | dre                             | ○●○                             | ○●○                             | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |
|      | cmi                             | ○                               | ○                               | ○●○                                 | ○●○                    | ○●○                               | ○●○                               |         |         |

Previously annotated miRNAs (mirBase 18) are depicted by filled circles, newly found putative miRNAs are shown as empty circles. Dashed lines separate different lineages: Primates + *Tupaia* (1), Glires (2), Euarchontoglires (1+2), Laurasiatheria (3), Afrotheria (4), Xenarthra (5), Eutheria (1-5), Metatheria (6), Saurospsida (7), Teleostei (8). The two paralogous sets of clusters are separated by the long-dashed line.



**Figure 1.** Estimated phylogenetic trees for the let-7 miRNA sequences. The two trees contain selected sequences from clusters A, C, and D (left) and E to J (right). The numbers at branches indicate the posterior support. Subtrees that are specific for teleosts (label-prefix “teleost”) or non-teleosts (no label-prefix) were collapsed to increase readability of the right tree, even if the subtrees are not complete.

Both the three mixed (A, C, and D) and the four homogeneous (E, F, G, and H) let-7 clusters clearly are the result of the vertebrate-specific (2R) genome duplications.

The situation is more complex in the five teleosts due to an extra round of genome duplication. The fish-specific genome duplication (FSGD) preceded the divergence of the teleosts.<sup>53,54</sup> Combining synteny information and sequence comparison allows to resolve the orthology relationships of the let-7 loci among the teleosts, see Table 2.

The correspondence of tetrapod and teleost let-7 clusters cannot be determined based on sequence similarity alone due to the short sequences and the large phylogenetic distances. For most loci, strong support comes again from synteny information. The teleost Aa locus, in particular, shares several flanking protein-coding genes with the human A locus, e.g., SMG9 and HAS1. We note, however, that the sequences of the teleost Aa and Ab loci are not recognizable as orthologs of the tetrapod A locus, while synteny and sequence data are largely consistent for the other loci. There is no support for the alternative explanation either, namely that the teleost clusters Aa and Ab derive really from an ancestral gnathostome B locus that has been lost completely in Tetrapoda, while the A locus has completely disappeared in teleosts. Taken

together, thus, the data suggest an ancestral state prior to the FSGD that closely matches the ancestral state in gnathostomes, see Figure 2. The only changes that can be attributed to the actinopterygian stem lineages are the loss of A-mir-100 and E-let-7-3. Surprisingly, the loci I and J are found in a loose association with the H and G clusters. In the case of G/I, this association is found in both paralogs, implying that the proximity of G and I loci preceded the FSGD. The G and I loci are also found on the same chromosomes, although separated by many megabases, in rat, dog, cow, sheep, and in saurospids.

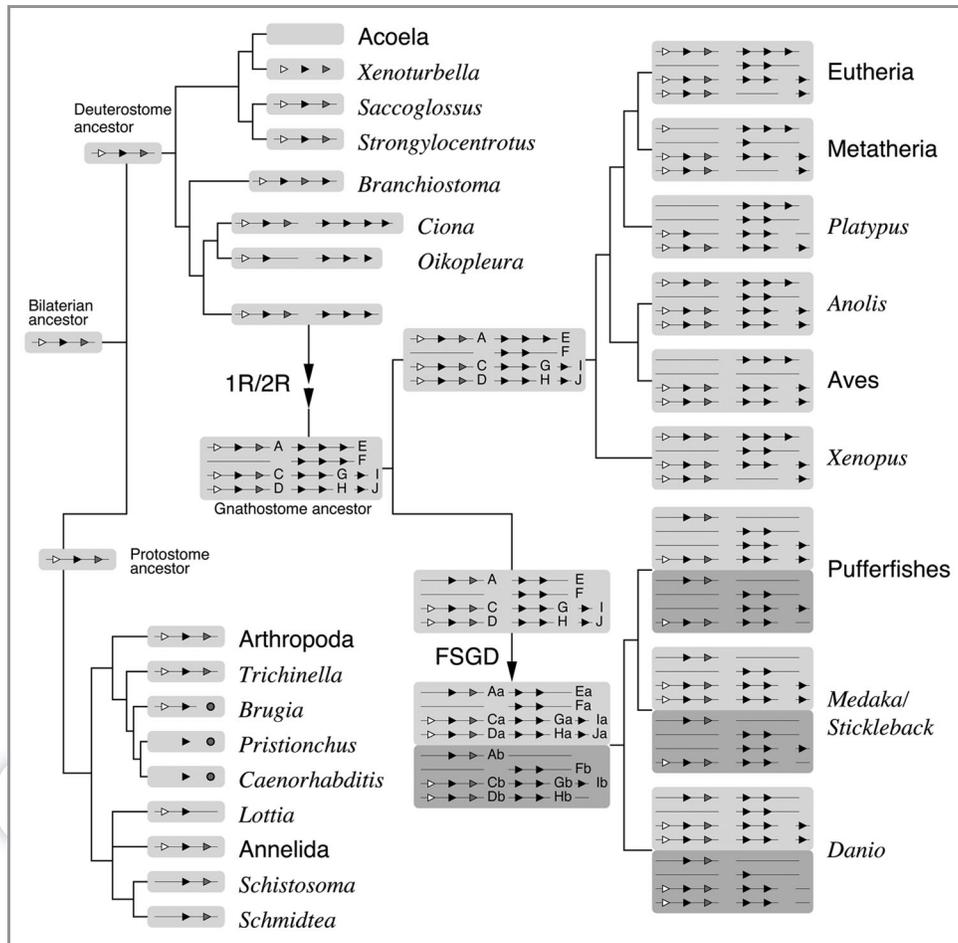
During gnathostome evolution, we observe several clade-specific loss events of entire clusters and individual microRNAs, cf. Figure 2. The most dramatic reductions occur in the wake of the FSGD with the complete loss of one copy of clusters E and J, the subsequent loss of the other copy of E in the percomorph lineage (pufferfishes, medaka, and stickleback), and the deletion of the Ca cluster in pufferfishes. Aves lack both the A and the F clusters, both of which are still present in the lizard genome. This could be due to the bird genome assemblies, which are, however, known to be incomplete in particular in their coverage of the micro-chromosomes.<sup>55</sup> Among mammals, only platypus features an H cluster, while this locus is lost in all Theria. Other

missing clusters in Eutheria affect in particular low coverage genomes and might be explained better by an incomplete assembly. A conspicuous pattern is the lack of the A cluster in the lemurs, however. A

**Table 2.** Correspondence of let-7 loci in the teleost genomes of *Danio rerio* (dre), *Oryzias latipes* (ola), *Gasterosteus aculeatus* (gac), *Takifugu rubripes* (tru), and *Tetraodon nigroviridis* (tni)

| loc. | dre                | ola               | gac                 | tru    | tni                |
|------|--------------------|-------------------|---------------------|--------|--------------------|
| Aa   | 16 <sup>27M</sup>  | 16 <sup>12M</sup> | XX <sup>8M</sup>    | s.22   | 8 <sup>6M</sup>    |
| Ab   | 19 <sup>10M</sup>  | s.1995            | X <sup>5.6M</sup>   | s.37   | 21-rnd             |
| Ca   | 10 <sup>39M</sup>  | 14 <sup>1M</sup>  | VII <sup>13M</sup>  | —      | —                  |
| Cb   | 15 <sup>29M</sup>  | —                 | —                   | —      | —                  |
| Da   | 15 <sup>20M</sup>  | 13 <sup>2M</sup>  | —                   | s.144  | 16 <sup>8M</sup>   |
| Db   | 5 <sup>31M</sup>   | 14 <sup>16M</sup> | VII <sup>18M</sup>  | s.6    | 7 <sup>6M</sup>    |
| Ea   | 11 <sup>28M</sup>  | —                 | —                   | —      | —                  |
| Ha   | 6 <sup>54M</sup>   | 5 <sup>12M</sup>  | XVII <sup>12M</sup> | s.7484 | 11 <sup>9M</sup>   |
| Ja   | 6 <sup>41M</sup>   | 5 <sup>7M</sup>   | c.7697              | s.56   | 11 <sup>5M</sup>   |
| Hb   | 23 <sup>5M</sup>   | 7 <sup>13M</sup>  | XII <sup>12M</sup>  | s.93   | 9 <sup>4.6M</sup>  |
| Fa   | 23 <sup>28M</sup>  | 7 <sup>12M</sup>  | XII <sup>11M</sup>  | s.66   | 9 <sup>3.6M</sup>  |
| Fb   | 23 <sup>18M</sup>  | 7 <sup>26M</sup>  | XII <sup>13M</sup>  | s.2159 | 9 <sup>10M</sup>   |
| Ga   | 25 <sup>2M</sup>   | 6 <sup>6M</sup>   | XIX <sup>4.8M</sup> | s.2    | 13 <sup>9.2M</sup> |
| Ia   | 25 <sup>1.6M</sup> | 6 <sup>5.8M</sup> | XIX <sup>4.5M</sup> | s.2    | 13 <sup>9.0M</sup> |
| Gb   | 4 <sup>17M</sup>   | 23 <sup>6M</sup>  | IV <sup>19M</sup>   | s.177  | 19 <sup>5.3M</sup> |
| Ib   | —                  | s.1942            | IV <sup>20M</sup>   | s.177  | 19 <sup>5.4M</sup> |

Synteny was determined from the pairwise alignment nets<sup>19</sup> provided through the UCSC genome browser. Note that the genomic coordinates of each loci are abbreviated by the chromosome or scaffold number and their position on megabase scale in superscript.



**Figure 2.** Putative evolutionary history of the let-7 microRNA clusters across Bilateria. A white triangle represents a mir-100, a gray triangle represents mir-125 microRNAs while let-7 sequences are depicted by black triangles. Annotated lin-4 microRNAs are shown with circles. 1R/2R denote two rounds of whole genome duplications, whereas FSGD labels the additional teleost-specific genome duplication. The duplicate clusters in teleosts are highlighted by different shading. Entire lineages are written in bold, genera are written in italic. Dispersed, highly derived let-7 and mir-125 paralogs in platyhelminthes and in *Brugia* are not shown.

gain of new let-7 loci is observed only in primates. The L locus appears in Haplorhini (tarsier and monkeys), while the K locus is present in Catarrhini (old world monkeys) only. We found evidence of expression of the miR sequence of K and L loci in miRNA-seq data of Human and Rhesus macaque brain samples<sup>56</sup> as well as in small RNA-seq data of the ENCODE cell lines.<sup>57</sup>

The ancestral let-7 clusters were apparently tightly linked as one would expect from a polycistronic primary transcript. However, some of these distances in mixed clusters substantially increase in tetrapods, namely, D-mir-100/D-let-7, C-let-7/C-mir-125, and D-let-7/D-mir-125, see also **Figure 3**. In both cases, the entire clusters are contained in the introns of non-coding

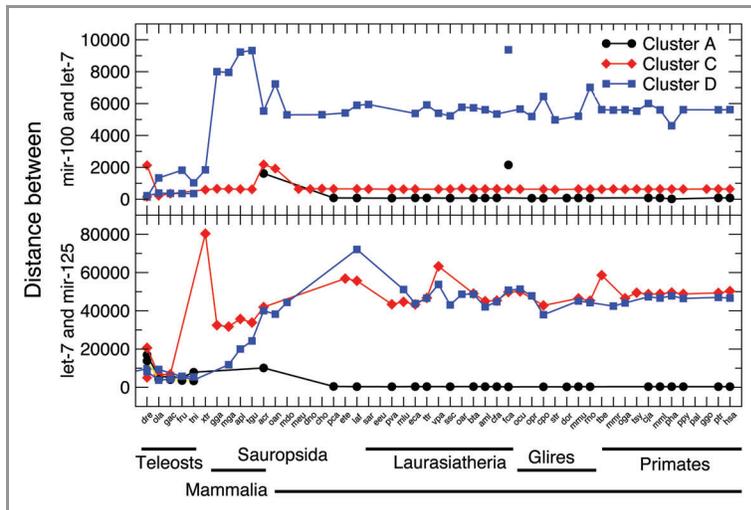
primary precursors, in human known as LINC00478 (C-cluster) and MIR100HG (D-cluster), respectively. Cluster F is expressed from an intron of the coding HUWE1 gene throughout and locus J is conserved within an intron of the coding WDR82 gene. Cluster G, in contrast, is exonic, located in the 3' exon of the non-coding host gene MIRLET7BHG. Loci E and I are associated with clusters of unspliced ESTs, the expression of cluster A cannot be resolved from currently available data.

Interestingly, human MIRLET7BHG harbors in one of its introns the additional annotated microRNA mir-3619 where evidence of expression was found in small RNA-seq data from embryonic stem cells.<sup>58</sup> This is an evolutionarily

young innovation, present only in old world monkeys.

By the use of family-wide covariance models of all let-7 families, clear evidence for the close relationships of let-7 microRNA of mixed clusters can be found, c.f. **Figure 4**. Furthermore, all let-7-1 and let-7-2 cluster appear to be closely related to each other corroborating their origin by genome duplications.

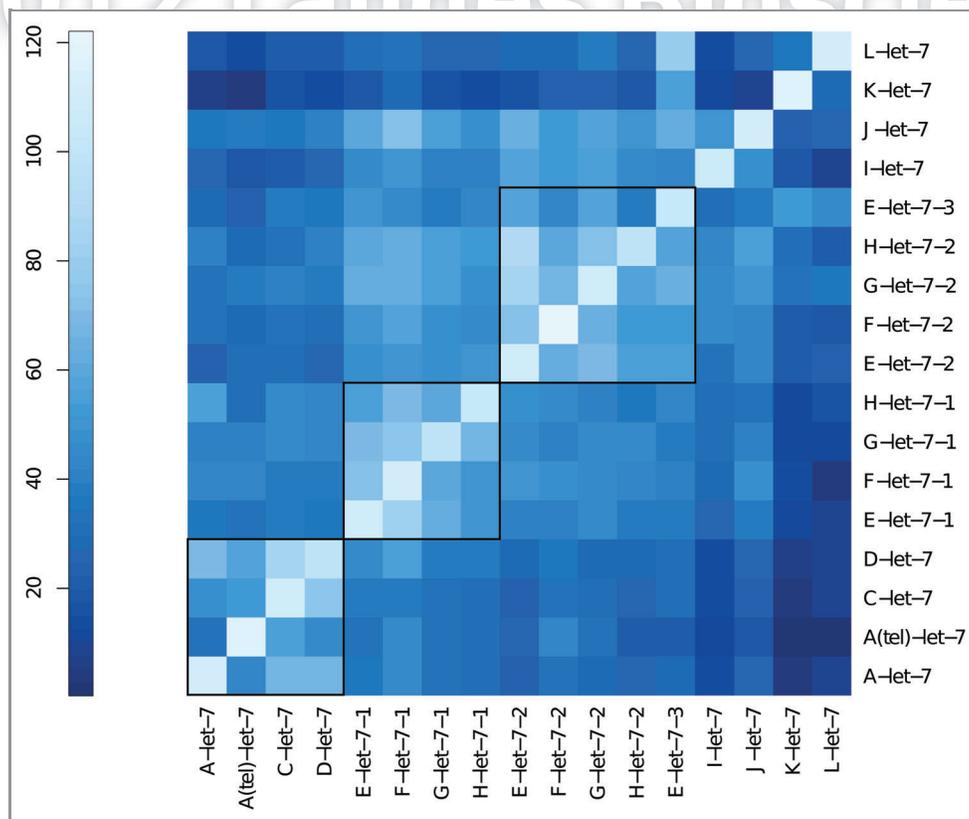
**Let-7 in Basal Deuterostomes.** Cyclostomia, lampreys and hagfishes, share at least one and possibly both rounds of the vertebrate genome duplication. Genomic data are solely available for the lamprey *Petromyzon marinus*. The miRBase lists 7 let-7, 3 mir-100, and a single mir-125,<sup>59</sup> not all of which can be recovered from the available genome assembly. The



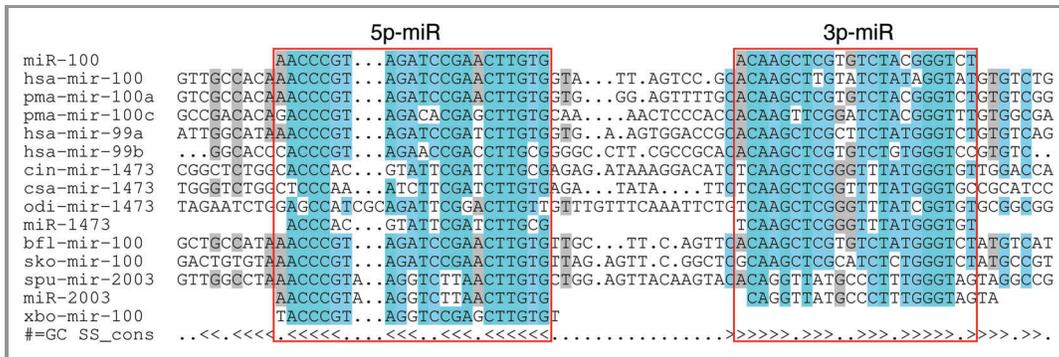
**Figure 3.** Distances between the pairs of adjacent pairs of microRNAs (mir-100/let-7 and let-7/mir-125) in the mixed clusters A, C, and D are conserved across Mammalia. Locus A is quite compact throughout the gnathostomes. Cluster D, in contrast, shows consistently large distances in tetrapods. In the cat *Felis catus*, distance outliers are owed to large gaps in the assembly within clusters.

pma-mir-100c sequence, however, is the reverse complement of pma-mir-100a.

The mir-100a and mir-125 genes are genomically linked and hence derive from the ancestral mixed cluster, although the corresponding let-7 expected to lie between them is missing. The presence of pma-mir-100b (without a linked mir-125 paralogs) serves as a witness of at least one round of genome duplication. Based on the similarities measured with the help of covariance models, we can also identify pma-let-7a-4 as descendant of the let-7 located in the ancestral mixed cluster. It seems to be the only lamprey let-7 miRNA originating from a mixed cluster. However, the assignment of the corresponding cluster is not possible since the sequence is not mappable to any available genome assembly. Among the homogeneous clusters, pma-let-7d and pma-let-7a-3 form a cluster; for the other loci the



**Figure 4.** Heatmap illustrating the structural and sequence similarity of all let-7 families. Therefore, all let-7 microRNA sequences of each family were scored against each family-wide covariance model. The average bitscore of covariance model (row) vs. let-7 family (column) is visualized in a color gradient. The standard deviations of these bitscores is always below 15 with a median bitscore standard deviation of 3.4. Due to the sequence divergence of cluster A in tetrapods and Aa/Ab in teleost fishes, both lineages were analyzed separately.



**Figure 5.** Highly derived mir-100 paralogs in basal deuterostomes. Although the seed regions also contain substitutions, the homology is clearly visible.

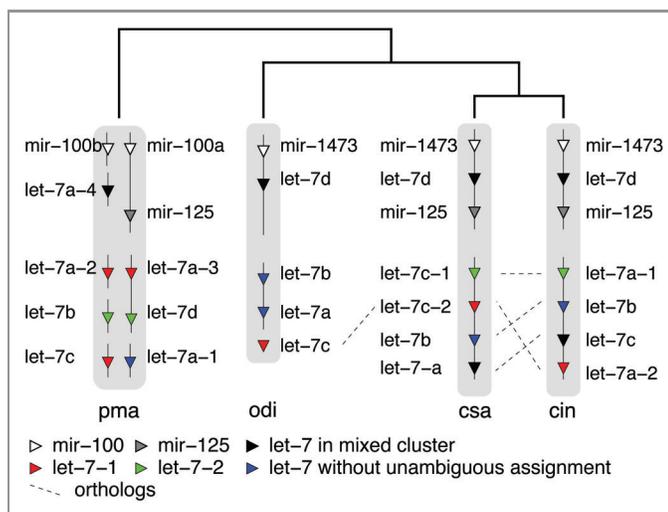
genome assembly is too fragmented to be informative. By applying Infernal to these miRNAs, strong evidence is obtained that *P. marinus* contains at least 2 homogeneous clusters comprising both a let-7-1 and a let-7-2 paralogs. One of the two remaining sequences, pma-let-7c, shows also obvious characteristics of a let-7-1 subfamily, whereas the other one, pma-let-7a-1, reveals no clear features to make a precise assignment of its origin. Sequencing of short RNAs also shows the presence of, presumably, multiple copies of mir-100, mir-125, and let-7 in the genome of the hagfish *Myxine glutinosa*.<sup>59</sup>

In *Ciona intestinalis*, an intron of an EST cluster that is homologous to the HUWE1 protein harbours a closely spaced cluster of four copies of let-7. On a different chromosome, there is a single mixed cluster consisting of mir-1473, let-7d, and mir-125. A very similar arrangement is found in *Ciona savignyi*. Upon closer inspection, cin-mir-1473 and csa-mir-1473 are clearly homologs of mir-100 revealing an ancestral mixed cluster, see Figure 5. In *Oikopleura dioica*, one locus harbours a let-7 and a mir-1473/mir-100 ortholog,<sup>4</sup> a second locus consists of two let-7 paralogs, and a final copy of let-7 is found on a third scaffold. The cin-let-7a-1 and cin-let-7a-2, as well as their counterparts csa-let-7c-1 and csa-let-7c-2, appear to be homologs of the let-7-2 and let-7-1 subfamily, respectively. In *C. intestinalis*, let-7a-1 is the first miRNA in the homogeneous let-7 cluster while let-7a-2 is located at the end. On the other hand, both *C. savignyi* sequences are located at the beginning. In *O. dioica*, however, solely odi-let-7c located on scaffold 10

appears to be assignable to the let-7-1 subfamily whereas the corresponding member of the let-7-2 subfamily is missing. Both cin-let-7c and its ortholog csa-let-7a might be copies of the let-7 miRNA of the mixed cluster. Unfortunately, the other orthologous miRNAs cin-let-7b and csa-let-7b cannot clearly be assigned to any let-7 subfamily by the use of covariance models, although their 5p-miR appears to be closely related to let-7 miRNAs originating from mixed clusters. Nevertheless, both remaining *O. dioica* sequences, namely odi-let-7a and odi-let-7b, are neither unambiguously assignable to any ascidian let-7 miRNA nor to any other subfamily. The relationships of the let-7 loci in basal deuterostomes are summarized in Figure 6.

In the lancelet *Branchiostoma floridae*, there is a second copy of let-7 located at the 3' end of the canonical cluster.<sup>60</sup> The two bfl-let-7 precursor sequences differ by only 4 point mutations, and hence are probably the results of a lineage-specific duplication. The ancestral cluster mir-100/let-7/mir-125 is also present in ambulacrarians, i.e., the acorn worm (hemichordata) and the sea urchin (echinodermata). In the latter, the mir-100 ortholog is rather diverged and recorded as spu-mir-2003 in miRBase.

The most basal clade of Deuterostomia, recently termed Xenacoelomorpha,<sup>61</sup> is composed of Xenoturbellida and Acoelomorpha. For *Xenoturbella bocki*, mature sequences of mir-125, let-7, and mir-100 have been reported.<sup>61</sup> In contrast, no



**Figure 6.** Relationships of let-7 loci in the basal deuterostomes *P. marinus* (pma), *O. dioica* (odi), *C. savignyi* (csa), and *C. intestinalis* (cin). Assignments to certain let-7 subfamilies were made with the use of Infernal. For details, see text.

evidence for any of the three microRNA families were found in microRNA libraries of *Hofstenia miamia*<sup>61</sup> and *Symsagittifera roscoffensis*.<sup>28,62</sup> A survey of several acoels by northern blot also returned a negative result.<sup>11</sup>

**Antisense microRNAs and cluster extensions.** MicroRNA precursors are very stable hairpin structures so that their 3' and 5' halves are close to being reverse complements. An antisense transcript therefore will in general also give rise to pre-mir-like hairpin structure. Indeed, functional antisense microRNAs have been reported in the literature for several loci, see e.g. Bender.<sup>63</sup> It does not come as a surprise, therefore, that antisense microRNAs have been found in deep sequencing data for some of the let-7 loci, cf. Table 3. In this survey, however, microRNAs that are located antisense to miRBase annotated mir-100, mir-125, or let-7 microRNAs were not further investigated.

Hairpins of about the size of pre-microRNAs are among the most frequent secondary structure motifs. This provides a likely mechanism for the innovation of new microRNAs.<sup>6</sup> The current version of miRBase reports several cases of additional microRNAs that emerged within or closely adjacent to a let-7 cluster. The best conserved example is hsa-mir-4763<sup>64</sup> located in the G cluster between hsa-let-7a-3 and hsa-let-7b. The human miR sequence is fairly well conserved in primates and to some extent in eutheria, see the corresponding alignment in the supplement material. However, there was no clear evidence of expression of the annotated hsa-mir-4763 in brain samples<sup>56</sup> or ENCODE cell lines.<sup>57</sup> Furthermore, conserved orthologs in *Macaca mulatta*

and *Canis familiaris* revealed no expression in miRNA-seq data from rhesus macaque brain<sup>56</sup> or in small RNA-seq data from domestic dog lymphocytes,<sup>65</sup> respectively. The cow microRNA bta-mir-2443, that is also located in cluster G between both let-7 microRNAs, is not an ortholog of the hsa-mir-4763 sequence. There is evidence for its expression in small RNA libraries of bovine kidney cells.<sup>66</sup> An interesting finding is the only detectable ortholog found in the dolphin *Tursiops truncatus*. This microRNA is located upstream of the mir-4763 ortholog and of the G-let-7-2 sequence, while the corresponding let-7-1 sequences of cluster G is missing. In *Bombyx mori*, bmo-mir-2795 is inserted between let-7 and mir-100.<sup>67</sup> A search of the NCBI databases did not reveal homologs in other insects. In the zebra finch, finally, tgu-mir-2987 is found about 3.1 kb downstream of the E cluster. Its sequence is not conserved in other bird genomes.

## Discussion

The association of mir-100, mir-125, and let-7 with its key conserved function in developmental timing<sup>32</sup> is one of the evolutionarily most ancient systems of microRNA-based regulation. The ancestral cluster of these three microRNAs dates back to the advent of Bilateria. In fact, only mir-100 and mir-10 date back further and are common to Eubilateria.<sup>3-5</sup> The evolution of the let-7 cluster in Protostomia is characterized mostly by partial losses and only occasional gene duplications (e.g., mir-100 in *Brugia* and mir-125 in platyhelminthes). In contrast, early chordates have acquired a second let-7 locus that subsequently expanded by

tandem duplication. The vertebrate-specific genome duplications expanded this system to a large number of paralogous loci. The retention rate of these paralogs is rather high, with up to 20 let-7 cluster microRNAs present in extant tetrapods, compared with the ancestral 24 microRNAs that are inferred from two rounds of duplications of the two chordate clusters. This is comparable with the fate of important transcriptional regulators such as the HOX gene clusters,<sup>68,69</sup> while the redundancy generated by genome duplications is nearly completely resolved, e.g., for metabolic enzymes.

The detailed analysis of the let-7 family also shows that microRNAs are always as conserved as one might expect. Beyond loss events, we also found highly derived paralogs that by combination of synteny and sequence similarity are unambiguously recognizable as homologs. The best examples are the homology of lin-4 and mir-125 in nematodes and the mir-100 paralogs mir-1473 (tunicates) and mir-2003 (echinoderms). This observation suggests that undocumented homologies are present also among other annotated microRNA families and it has an impact on the use of microRNAs as a phylogenetic marker as unrecognized derived microRNA families can be misinterpreted as the ancestral state in which the microRNA family has not yet emerged.

The naming convention of miRBase for paralogous microRNAs has turned out to be a major technical inconvenience for the present study. True orthologs (as determined by both synteny and sequence comparison of the complete precursor sequences) not infrequently have different names in different species. Even worse, paralogous copies may have the same name. It would be desirable, therefore, to rethink the naming schemes to convey information on the genomic location. For vault RNAs, which also form multiple clusters in mammalian genomes genes, names that make the cluster membership explicit were recently adopted by the HGNC.<sup>70</sup>

## Acknowledgments

This work is based on the results of a bioinformatics computer lab course in the Winter Semester 2010/2011. The following

**Table 3.** Antisense microRNAs associated with let-7 loci listed in miRBase v.18

| Species  | loc. | sense          | antisense     | Ref. |
|----------|------|----------------|---------------|------|
| rat      | D    | rno-let-7a-2   | rno-mir-3596a | 68   |
| rat      | A    | rno-let-7e     | rno-mir-3596c | 68   |
| rat      | E-1  | rno-let-7f-1   | rno-mir-3596d | 68   |
| rat      | E-3  | rno-let-7d     | rno-mir-3596b | 68   |
| rat      | C    | rno-mir-125b-2 | rno-mir-3588  | 68   |
| cow      | G-2  | bta-let-7b     | bta-mir-3596  | 69   |
| lancelet | —    | bfl-let-7a-1   | bfl-let-7b    | 70   |
| lancelet | —    | bfl-mir-125a   | bfl-mir-125b  | 70   |

students contributed their preliminary analysis to this work: Denise Aumer, Simon Bin, Peter Buske, Jörg Dreier, Katharina Eichler, Martin Fischer, Belinda Kahnt, Rebecca Kirsch, Martin Krüger,

Vera Lede, Sandra Treffkorn, Sigrid Uxa, Sarah Witzsche. This work was funded in part by the European Regional Development Fund (ERDF) and by the Free State of Saxony within the framework

of LIFE – Leipzig Research Center for Civilization Diseases (CO), a PhD stipend funded by the European Social Fund (AW), and the European FP-7 project QUANTOMICS (PFS,SB, no. 222664).

## References

- Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 2005; 120:21-4; PMID:15652478; <http://dx.doi.org/10.1016/j.cell.2004.12.031>
- Sempere LF, Cole CN, McPeck MA, Peterson KJ. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol* 2006; 306:575-88; PMID:16838302; <http://dx.doi.org/10.1002/jez.b.21118>
- Heimberg AM, Sempere LF, Moy VN, Donoghue PCJ, Peterson KJ. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* 2008; 105:2946-50; PMID:18287013; <http://dx.doi.org/10.1073/pnas.0712259105>
- Fu X, Adamski M, Thompson EM. Altered miRNA repertoire in the simplified chordate, *Oikopleura dioica*. *Mol Biol Evol* 2008; 25:1067-80; PMID:18339653; <http://dx.doi.org/10.1093/molbev/msn060>
- Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, et al. The expansion of the metazoan microRNA repertoire. *BMC Genomics* 2006; 7:15; PMID:16438707; <http://dx.doi.org/10.1186/1471-2164-7-25>
- Tanzer A, Stadler PF. Molecular evolution of a microRNA cluster. *J Mol Biol* 2004; 339:327-35; PMID:15136036; <http://dx.doi.org/10.1016/j.jmb.2004.03.065>
- Tanzer A, Stadler PF. Evolution of microRNAs. In SY Ying, editor, *MicroRNA Protocols*, volume 342 of *Methods in Molecular Biology*. Humana Press, Totowa, NJ, 2006, 335–350.
- Ambros V. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. *Cell* 1989; 57:49-57; PMID:2702689; [http://dx.doi.org/10.1016/0092-8674\(89\)90171-2](http://dx.doi.org/10.1016/0092-8674(89)90171-2)
- Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 2001; 294:858-62; PMID:11679671; <http://dx.doi.org/10.1126/science.1065062>
- Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 2000; 408:86-9; PMID:11081512; <http://dx.doi.org/10.1038/35040556>
- Pasquinelli AE, McCoy A, Jiménez E, Saló E, Ruvkun G, Martindale MQ, et al. Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: a role in life history evolution? *Evol Dev* 2003; 5:372-8; PMID:12823453; <http://dx.doi.org/10.1046/j.1525-142X.2003.03044.x>
- Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, et al. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 2000; 403:901-6; PMID:10706289; <http://dx.doi.org/10.1038/35002607>
- Roush S, Slack FJ. The let-7 family of microRNAs. *Trends Cell Biol* 2008; 18:505-16; PMID:18774294; <http://dx.doi.org/10.1016/j.tcb.2008.07.007>
- Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res* 2004; 32(Database issue):D109-11; PMID:14681370; <http://dx.doi.org/10.1093/nar/gkh023>
- Bompfünewerer AF, Flamm C, Fried C, Fritsch G, Hofacker IL, Lehmann J, et al. Evolutionary patterns of non-coding RNAs. *Theory Biosci* 2005; 123:301-69; PMID:18202870; <http://dx.doi.org/10.1016/j.thbio.2005.01.002>
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; 215:403-10; PMID:2231712
- Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, et al. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res* 2009; 37:1602-15; PMID:19151082; <http://dx.doi.org/10.1093/nar/gkn1084>
- Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009; 25:1335-7; PMID:19307242; <http://dx.doi.org/10.1093/bioinformatics/btp157>
- Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 2003; 100:11484-9; PMID:14500911; <http://dx.doi.org/10.1073/pnas.1932072100>
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer SL, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994; 125:167-88; <http://dx.doi.org/10.1007/BF00818163>
- Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 2008; 9:474; PMID:19014431; <http://dx.doi.org/10.1186/1471-2105-9-474>
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; 23:2947-8; PMID:17846036; <http://dx.doi.org/10.1093/bioinformatics/btm404>
- Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006; 23:254-67; PMID:16221896; <http://dx.doi.org/10.1093/molbev/msj030>
- Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003; 19:1572-4; PMID:12912839; <http://dx.doi.org/10.1093/bioinformatics/btg180>
- Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* 2008; 25:1253-6; PMID:18397919; <http://dx.doi.org/10.1093/molbev/msn083>
- Huson DH, Richter DC, Rausch C, DeZulian T, Franz M, Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 2007; 8:460; PMID:18034891; <http://dx.doi.org/10.1186/1471-2105-8-460>
- Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 2008; 455:1193-7; PMID:18830242; <http://dx.doi.org/10.1038/nature07415>
- Sempere LF, Martinez P, Cole C, Bagaña J, Peterson KJ. Phylogenetic distribution of microRNAs supports the basal position of acel flatworms and the polyphyly of Platyhelminthes. *Evol Dev* 2007; 9:409-15; PMID:17845513; <http://dx.doi.org/10.1111/j.1525-142X.2007.00180.x>
- Ball EE, Miller DJ. Phylogeny: the continuing classificatory conundrum of chaetognaths. *Curr Biol* 2006; 16:R593-6; PMID:16890517; <http://dx.doi.org/10.1016/j.cub.2006.07.006>
- Harzsch S, Müller CH. A new look at the ventral nerve centre of Sagitta: implications for the phylogenetic position of Chaetognatha (arrow worms) and the evolution of the bilaterian nervous system. *Front Zool* 2007; 4:14; PMID:17511857; <http://dx.doi.org/10.1186/1742-9994-4-14>
- Paps J, Bagaña J, Riutort M. Bilaterian phylogeny: a broad sampling of 13 nuclear genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal acelomorpha. *Mol Biol Evol* 2009; 26:2397-406; PMID:19107204; <http://dx.doi.org/10.1093/molbev/msp150>
- Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, et al. Ancient animal microRNAs and the evolution of tissue identity. *Nature* 2010; 463:1084-8; PMID:20118916; <http://dx.doi.org/10.1038/nature08744>
- Xue X, Sun J, Zhang Q, Wang Z, Huang Y, Pan W. Identification and characterization of novel microRNAs from *Schistosoma japonicum*. *PLoS One* 2008; 3:e4034; PMID:19107204; <http://dx.doi.org/10.1371/journal.pone.0004034>
- Huang J, Hao P, Chen H, Hu W, Yan Q, Liu F, et al. Genome-wide identification of *Schistosoma japonicum* microRNAs using a deep-sequencing approach. *PLoS One* 2009; 4:e8206; PMID:19997615; <http://dx.doi.org/10.1371/journal.pone.0008206>
- Wang Z, Xue X, Sun J, Luo R, Xu X, Jiang Y, et al. An "in-depth" description of the small non-coding RNA population of *Schistosoma japonicum* schistosomulum. *PLoS Negl Trop Dis* 2010; 4:e596; PMID:20161724; <http://dx.doi.org/10.1371/journal.pntd.0000596>
- Palakodeti D, Smielewska M, Graveley BR. MicroRNAs from the Planarian *Schmidtea mediterranea*: a model system for stem cell biology. *RNA* 2006; 12:1640-9; PMID:16849698; <http://dx.doi.org/10.1261/rna.117206>
- Friedländer MR, Adamidi C, Han T, Lebedeva S, Isenbarger TA, Hirst M, et al. High-resolution profiling and discovery of planarian small RNAs. *Proc Natl Acad Sci U S A* 2009; 106:11546-51; PMID:19564616; <http://dx.doi.org/10.1073/pnas.0905222106>
- Lu YC, Smielewska M, Palakodeti D, Lovci MT, Aigner S, Yeo GW, et al. Deep sequencing identifies new and regulated microRNAs in *Schmidtea mediterranea*. *RNA* 2009; 15:1483-91; PMID:19553344; <http://dx.doi.org/10.1261/rna.1702009>
- Copeland CS, Marz M, Rose D, Hertel J, Brindley PJ, Santana CB, et al. Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. *BMC Genomics* 2009; 10:464; PMID:19814823; <http://dx.doi.org/10.1186/1471-2164-10-464>
- Sokol NS, Xu P, Jan YN, Ambros V. Drosophila let-7 microRNA is required for remodeling of the neuromusculature during metamorphosis. *Genes Dev* 2008; 22:1591-6; PMID:18559475; <http://dx.doi.org/10.1101/gad.1671708>

41. de Wit E, Linsen SE, Cuppen E, Berezikov E. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res* 2009; 19:2064-74; PMID:19755563; <http://dx.doi.org/10.1101/gr.093781.109>
42. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 2006; 127:1193-207; PMID:17174894; <http://dx.doi.org/10.1016/j.cell.2006.10.040>
43. Poole CB, Davis PJ, Jin J, McReynolds LA. Cloning and bioinformatic identification of small RNAs in the filarial nematode, *Brugia malayi*. *Mol Biochem Parasitol* 2010; 169:87-94; PMID:19874857; <http://dx.doi.org/10.1016/j.molbiopara.2009.10.004>
44. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 1993; 75:843-54; PMID:8252621; [http://dx.doi.org/10.1016/0092-8674\(93\)90529-Y](http://dx.doi.org/10.1016/0092-8674(93)90529-Y)
45. Caygill EE, Johnston LA. Temporal regulation of metamorphic processes in *Drosophila* by the *let-7* and *miR-125* heterochronic microRNAs. *Curr Biol* 2008; 18:943-50; PMID:18571409; <http://dx.doi.org/10.1016/j.cub.2008.06.020>
46. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. Identification of tissue-specific microRNAs from mouse. *Curr Biol* 2002; 12:735-9; PMID:12007417; [http://dx.doi.org/10.1016/S0960-9822\(02\)00809-6](http://dx.doi.org/10.1016/S0960-9822(02)00809-6)
47. Tyler DM, Okamura K, Chung WJ, Hagen JW, Berezikov E, Hannon GJ, et al. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev* 2008; 22:26-36; PMID:18172163; <http://dx.doi.org/10.1101/gad.1615208>
48. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004; 432:695-716; PMID:15592404; <http://dx.doi.org/10.1038/nature03154>
49. Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, et al. Survey sequencing and comparative analysis of the elephant shark (*Callorhynchus milii*) genome. *PLoS Biol* 2007; 5:e101; PMID:17407382; <http://dx.doi.org/10.1371/journal.pbio.0050101>
50. Ohno S. *Evolution by gene duplication*. Berlin: Springer-Verlag, 1970.
51. Meyer A, Schart M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol* 1999; 11:699-704; PMID:10600714; [http://dx.doi.org/10.1016/S0955-0674\(99\)00039-3](http://dx.doi.org/10.1016/S0955-0674(99)00039-3)
52. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 2005; 3:e314; PMID:16128622; <http://dx.doi.org/10.1371/journal.pbio.0030314>
53. Hoegg S, Brinkmann H, Taylor JS, Meyer A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* 2004; 59:190-203; PMID:15486693; <http://dx.doi.org/10.1007/s00239-004-2613-z>
54. Crow KD, Stadler PF, Lynch VJ, Amemiya CT, Wagner GP. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol Biol Evol* 2006; 23:121-36; PMID:16162861; <http://dx.doi.org/10.1093/molbev/msj020>
55. Dodgson JB, Delany ME, Cheng HH. Poultry genome sequences: progress and outstanding challenges. *Cytogenet Genome Res* 2011; 134:19-26; PMID:21335957; <http://dx.doi.org/10.1159/000324413>
56. Somel M, Guo S, Fu N, Yan Z, Hu HY, Xu Y, et al. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res* 2010; 20:1207-18; PMID:20647238; <http://dx.doi.org/10.1101/gr.106849.110>
57. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. ENCODE Project Consortium, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute & Children's Hospital Oakland Research Institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; 447:799-816; PMID:17571346; <http://dx.doi.org/10.1038/nature05874>
58. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 2008; 18:610-21; PMID:18285502; <http://dx.doi.org/10.1101/gr.7179508>
59. Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PC, Peterson KJ. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci U S A* 2010; 107:19379-83; PMID:20959416; <http://dx.doi.org/10.1073/pnas.1010350107>
60. Campo-Paysaa F, Sémon M, Cameron RA, Peterson KJ, Schubert M. microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol Dev* 2011; 13:15-27; PMID:21210939; <http://dx.doi.org/10.1111/j.1525-142X.2010.00452.x>
61. Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, et al. Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* 2011; 470:255-8; PMID:21307940; <http://dx.doi.org/10.1038/nature09676>
62. Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, et al. The deep evolution of metazoan microRNAs. *Evol Dev* 2009; 11:50-68; PMID:19196333; <http://dx.doi.org/10.1111/j.1525-142X.2008.00302.x>
63. Bender W. MicroRNAs in the *Drosophila* bithorax complex. *Genes Dev* 2008; 22:14-9; PMID:18172161; <http://dx.doi.org/10.1101/gad.1614208>
64. Persson H, Kvist A, Rego N, Staaf J, Vallon-Christersson J, Luts L, et al. Identification of new microRNAs in paired normal and tumor breast tissue suggests a dual role for the ERBB2/Her2 gene. *Cancer Res* 2011; 71:78-86; PMID:21199797; <http://dx.doi.org/10.1158/0008-5472.CAN-10-1869>
65. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008; 26:407-15; PMID:18392026; <http://dx.doi.org/10.1038/nbt1394>
66. Glazov EA, Kongsuwan K, Assavalapsakul W, Horwood PF, Mitter N, Mahony TJ. Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. *PLoS One* 2009; 4:e6349; PMID:19633723; <http://dx.doi.org/10.1371/journal.pone.0006349>
67. Liu S, Li D, Li Q, Zhao P, Xiang Z, Xia Q. MicroRNAs of *Bombyx mori* identified by Solexa sequencing. *BMC Genomics* 2010; 11:148; PMID:20199675; <http://dx.doi.org/10.1186/1471-2164-11-148>
68. Prohaska SJ, Fried C, Amemiya CT, Ruddle FH, Wagner GP, Stadler PF. The shark HoxN cluster is homologous to the human HoxD cluster. *J Mol Evol* 2004; 58:212-7; PMID:15042342; <http://dx.doi.org/10.1007/s00239-003-2545-z>
69. Amemiya CT, Powers TP, Prohaska SJ, Grimwood J, Schmutz J, Dickson M, et al. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc Natl Acad Sci U S A* 2010; 107:3622-7; PMID:20139301; <http://dx.doi.org/10.1073/pnas.0914312107>
70. Stadler PF, Chen JJ, Hacker Müller J, Hoffmann S, Horn F, Khaitovich P, et al. Evolution of vault RNAs. *Mol Biol Evol* 2009; 26:1975-91; PMID:19491402; <http://dx.doi.org/10.1093/molbev/msp112>